8 SUPPLEMENTARY MATERIAL

8.1 The gradient of Eq. 2 and 3 with the standard normal prior

The gradient of $\widetilde{\mathcal{L}}(\phi,\theta,\lambda)$ in (2) and (3) for the standard normal prior with respect to a single weight ϕ_i for a single data point $\mathbf{x}$ is the following:

$$\frac{\partial}{\partial\phi_i}\widetilde{\mathcal{L}}(\mathbf{x};\phi,\theta,\lambda) = \frac{1}{L}\sum_{l=1}^{L}\Big[\frac{1}{p_\theta(\mathbf{x}|\mathbf{z}_\phi^{(l)})}\ \frac{\partial}{\partial\mathbf{z}_\phi}p_\theta(\mathbf{x}|\mathbf{z}_\phi^{(l)})\ \frac{\partial}{\partial\phi_i}\mathbf{z}_\phi^{(l)} - \frac{1}{q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\frac{\partial}{\partial\phi_i}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}) + \tag{17}$$

$$+\frac{1}{p_\lambda(\mathbf{z}_\phi^{(l)})\ q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\Big(q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\frac{\partial}{\partial\mathbf{z}_\phi}p_\lambda(\mathbf{z}_\phi^{(l)}) - p_\lambda(\mathbf{z}_\phi^{(l)})\frac{\partial}{\partial\mathbf{z}_\phi}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\Big)\frac{\partial}{\partial\phi_i}\mathbf{z}_\phi^{(l)}\Big]. \tag{18}$$

8.2 The gradient of Eq. 2 and 3 with the VampPrior

The gradient of $\widetilde{\mathcal{L}}(\phi,\theta,\lambda)$ in (2) and (3) for the VampPrior with respect to a single weight ϕ_i for a single data point $\mathbf{x}$ is the following:

$$\frac{\partial}{\partial\phi_i}\widetilde{\mathcal{L}}(\mathbf{x};\phi,\theta,\lambda) = \frac{1}{L}\sum_{l=1}^{L}\Big[\frac{1}{p_\theta(\mathbf{x}|\mathbf{z}_\phi^{(l)})}\ \frac{\partial}{\partial\mathbf{z}_\phi}p_\theta(\mathbf{x}|\mathbf{z}_\phi^{(l)})\ \frac{\partial}{\partial\phi_i}\mathbf{z}_\phi^{(l)} + \tag{19}$$

$$+\frac{1}{K}\sum_{k=1}^{K}\Big\{\ \Big(\frac{q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\ \frac{\partial}{\partial\phi_i}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ \frac{\partial}{\partial\phi_i}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}{\frac{1}{K}\sum_{k=1}^{K}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\Big) + \tag{20}$$

$$+\Big(\frac{\big(q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\ \frac{\partial}{\partial\mathbf{z}_\phi}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ \frac{\partial}{\partial\mathbf{z}_\phi}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\big)\ \frac{\partial}{\partial\phi_i}\mathbf{z}_\phi^{(l)}}{\frac{1}{K}\sum_{k=1}^{K}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\Big)\Big\}\Big]. \tag{21}$$

8.3 Details on the gradient calculation in Eq. 20 and 21

Let us recall the objective function for single datapoint $\mathbf{x}_*$ using L Monte Carlo sample points:

$$\widetilde{\mathcal{L}}(\mathbf{x}_*;\phi,\theta,\lambda) = \frac{1}{L}\sum_{l=1}^{L}\Big[\ln p_\theta(\mathbf{x}_*|\mathbf{z}_\phi^{(l)})\Big] + \frac{1}{L}\sum_{l=1}^{L}\Big[\ln\frac{1}{K}\sum_{k=1}^{K}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - \ln q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}_*)\Big]. \tag{22}$$

We are interested in calculating gradient with respect to a single parameter ϕ_i. We can split the gradient into two parts:

$$\frac{\partial}{\partial\phi_i}\widetilde{\mathcal{L}}(\mathbf{x}_*;\phi,\theta,\lambda) = \frac{\partial}{\partial\phi_i}\underbrace{\frac{1}{L}\sum_{l=1}^{L}\Big[\ln p_\theta(\mathbf{x}_*|\mathbf{z}_\phi^{(l)})\Big]}_{(*)}$$

$$+\frac{\partial}{\partial\phi_i}\underbrace{\frac{1}{L}\sum_{l=1}^{L}\Big[\ln\frac{1}{K}\sum_{k=1}^{K}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - \ln q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}_*)\Big]}_{(**)} \tag{23}$$

Calculating the gradient separately for both $(*)$ and $(**)$ yields:

$$\begin{aligned}
\frac{\partial}{\partial \phi_i}(*) &= \frac{\partial}{\partial \phi_i} \frac{1}{L} \sum_{l=1}^{L} \Big[\ln p_\theta(\mathbf{x}_*|\mathbf{z}_\phi^{(l)}) \Big] \\
&= \frac{1}{L} \sum_{l=1}^{L} \frac{1}{p_\theta(\mathbf{x}_*|\mathbf{z}_\phi^{(l)})} \; \frac{\partial}{\partial \mathbf{z}_\phi} p_\theta(\mathbf{x}_*|\mathbf{z}_\phi^{(l)}) \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)}
\end{aligned} \tag{24}$$

$$\begin{aligned}
\frac{\partial}{\partial \phi_i}(**) &= \frac{\partial}{\partial \phi_i} \frac{1}{L} \sum_{l=1}^{L} \Big[\ln \frac{1}{K} \sum_{k=1}^{K} q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - \ln q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}_*) \Big] \\
&\text{[Short-hand notation: } q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}_*) \triangleq q_\phi^*, \quad q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) \triangleq q_\phi^k] \\
&= \frac{1}{L} \sum_{l=1}^{L} \Big[\frac{1}{\frac{1}{K}\sum_{k=1}^{K} q_\phi^k} \Big(\frac{\partial}{\partial \phi_i} \frac{1}{K} \sum_{k=1}^{K} q_\phi^k + \frac{\partial}{\partial \mathbf{z}_\phi} \big(\frac{1}{K} \sum_{k=1}^{K} q_\phi^k \big) \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big) + \\
&\quad - \frac{1}{q_\phi^*} \Big(\frac{\partial}{\partial \phi_i} q_\phi^* + \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^* \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big) \Big] \\
&= \frac{1}{L} \sum_{l=1}^{L} \Big[\frac{1}{\frac{1}{K}\sum_{k=1}^{K} q_\phi^k \; q_\phi^*} \Big(\frac{1}{K} \sum_{k=1}^{K} q_\phi^* \; \frac{\partial}{\partial \phi_i} q_\phi^k + \big(\frac{1}{K} \sum_{k=1}^{N} q_\phi^* \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^k \big) \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big) + \\
&\quad - \frac{1}{\frac{1}{K}\sum_{k=1}^{K} q_\phi^k \; q_\phi^*} \Big(\frac{1}{K} \sum_{k=1}^{K} q_\phi^k \; \frac{\partial}{\partial \phi_i} q_\phi^* + \frac{1}{K} \sum_{k=1}^{K} q_\phi^k \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^* \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big) \Big] \\
&= \frac{1}{L} \sum_{l=1}^{L} \Big[\frac{1}{\frac{1}{K}\sum_{k=1}^{K} q_\phi^k \; q_\phi^*} \; \frac{1}{K} \sum_{k=1}^{K} \Big\{ \Big(q_\phi^* \; \frac{\partial}{\partial \phi_i} q_\phi^k - q_\phi^k \; \frac{\partial}{\partial \phi_i} q_\phi^* \Big) + \\
&\quad + \Big(q_\phi^* \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^k - q_\phi^k \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^* \Big) \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big\} \Big]
\end{aligned} \tag{25}$$

For comparison, the gradient of $(**)$ for a prior $p_\lambda(\mathbf{z})$ that is independent of the variational posterior is the following:

$$\begin{aligned}
&\frac{\partial}{\partial \phi_i} \Big[\frac{1}{L} \sum_{l=1}^{L} \ln p_\lambda(\mathbf{z}_\phi^{(l)}) - \ln q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}_*) \Big] = \\
&\text{[Short-hand notation: } q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x}_*) \triangleq q_\phi^*, \qquad p_\lambda(\mathbf{z}_\phi^{(l)}) \triangleq p_\lambda] \\
&= \frac{1}{L} \sum_{l=1}^{L} \Big[\frac{1}{p_\lambda} \; \frac{\partial}{\partial \mathbf{z}_\phi} p_\lambda \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} - \frac{1}{q_\phi^*} \Big(\frac{\partial}{\partial \phi_i} q_\phi^* + \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^* \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big) \Big] \\
&= \frac{1}{L} \sum_{l=1}^{L} \Big[- \frac{1}{q_\phi^*} \frac{\partial}{\partial \phi_i} q_\phi^* + \frac{1}{p_\lambda \; q_\phi^*} \Big(q_\phi^* \; \frac{\partial}{\partial \mathbf{z}_\phi} p_\lambda - p_\lambda \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^* \Big) \; \frac{\partial}{\partial \phi_i} \mathbf{z}_\phi^{(l)} \Big]
\end{aligned} \tag{26}$$

We notice that in (25) if $q_\phi^* \approx q_\phi^k$ for some k, then the differences $(q_\phi^* \; \frac{\partial}{\partial \phi_i} q_\phi^k - q_\phi^k \; \frac{\partial}{\partial \phi_i} q_\phi^*)$ and $(q_\phi^* \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^k - q_\phi^k \; \frac{\partial}{\partial \mathbf{z}_\phi} q_\phi^*)$ are close to 0. Hence, the gradient points into an average of all dissimilar pseudo-inputs contrary to the gradient of the standard normal prior in (26) that pulls always towards $\mathbf{0}$. As a result, the encoder is trained so that to have large variance because it is attracted by all dissimilar points and due to this fact it assigns separate regions in the latent space to each datapoint. This effect should help the decoder to decode a hidden representation to an image much easier.

8.4 Details on experiments

All experiments were run on NVIDIA TITAN X Pascal. The code for our models is available online at `https://github.com/jmtomczak/vae_vampprior`.

8.4.1 Datasets used in the experiments

We carried out experiments using six image datasets: static and dynamic MNIST[3], OMNIGLOT[4] [20], Caltech 101 Silhouettes[5] [27], Frey Faces[6], and Histopathology patches [39]. Frey Faces contains images of size 28×20 and all other datasets contain 28×28 images. We distinguish between static MNIST with fixed binarizartion of images [21] and dynamic MNIST with dynamic binarization of data during training as in [34].

MNIST consists of hand-written digits split into 60,000 training datapoints and 10,000 test sample points. In order to perform model selection we put aside 10,000 images from the training set. We distinguish between static MNIST with fixed binarizartion of images[7] [21] and dynamic MNIST with dynamic binarization of data during training as in [34].

OMNIGLOT is a dataset containing 1,623 hand-written characters from 50 various alphabets. Each character is represented by about 20 images that makes the problem very challenging. The dataset is split into 24,345 training datapoints and 8,070 test images. We randomly pick 1,345 training examples for validation. During training we applied dynamic binarization of data similarly to dynamic MNIST.

Caltech 101 Silhouettes contains images representing silhouettes of 101 object classes. Each image is a filled, black polygon of an object on a white background. There are 4,100 training images, 2,264 validation datapoints and 2,307 test examples. The dataset is characterized by a small training sample size and many classes that makes the learning problem ambitious.

Frey Faces is a dataset of faces of a one person with different emotional expressions. The dataset consists of nearly 2,000 gray-scaled images. We randomly split them into 1,565 training images, 200 validation images and 200 test images. We repeated the experiment 3 times.

Histopathology is a dataset of histopathology patches of ten different biopsies containing cancer or anemia. The dataset consists of gray-scaled images divided into 6,800 training images, 2,000 validation images and 2,000 test images.

8.4.2 Additional results: Wall-clock times

Using our implementation, we have calculated wall-clock times for $K = 500$ (measured on MNIST) and $K = 1000$ (measured on OMNIGLOT). HVAE+VampPrior was about 1.4 times slower than the standard normal prior. ConvHVAE and PixelHVAE with the VampPrior resulted in the increased training times, respectively, by a factor of $\times 1.9/\times 2.1$ and $\times 1.4/\times 1.7$ ($K = 500/K = 1000$) comparing to the standard prior. We believe that this time burden is acceptable regarding the improved generative performance resulting from the usage of the VampPrior.

8.4.3 Additional results: Generations, reconstructions and histograms of log-likelihood

The generated images are presented in Figure 5. Images generated by HVAE ($L = 2$) + VampPrior are more realistic and sharper than the ones given by the vanilla VAE. The quality of images generated by convHVAE and PixelHVAE contain much more details and better reflect variations in data.

The reconstructions from test images are presented in Figure 6. At first glance the reconstructions of VAE and HVAE ($L = 2$) + VampPrior look similarly, however, our approach provides more details and the reconstructions are sharper. This is especially visible in the case of OMNIGLOT (middle row in Figure 6) where VAE is incapable to reconstruct small circles while our approach does in most cases. The application of convolutional networks further improves the quality of reconstructions by providing many tiny details. Interestingly, for the PixelHVAE we can notice some "fantasizing" during reconstructing images (*e.g.*, for OMNIGLOT). It means that the decoder was, to some extent, too flexible and disregarded some information included in the latent representation.

The histograms of the log-likelihood per test example are presented in Figure 7. We notice that all histograms

[3] `http://yann.lecun.com/exdb/mnist/`

[4] We used the pre-processed version of this dataset as in [6]: `https://github.com/yburda/iwae/blob/master/datasets/OMNIGLOT/chardata.mat`.

[5] We used the dataset with fixed split into training, validation and test sets: `https://people.cs.umass.edu/~marlin/data/caltech101_silhouettes_28_split1.mat`.

[6] `http://www.cs.nyu.edu/~roweis/data/frey_rawface.mat`

[7] `https://github.com/yburda/iwae/tree/master/datasets/BinaryMNIST`

characterize a heavy-tail indicating existence of examples that are hard to represent. However, taking a closer look at the histograms for HVAE ($L = 2$) + VampPrior and its convolutional-based version reveals that there are less hard examples comparing to the standard VAE. This effect is especially apparent for the convHVAE.

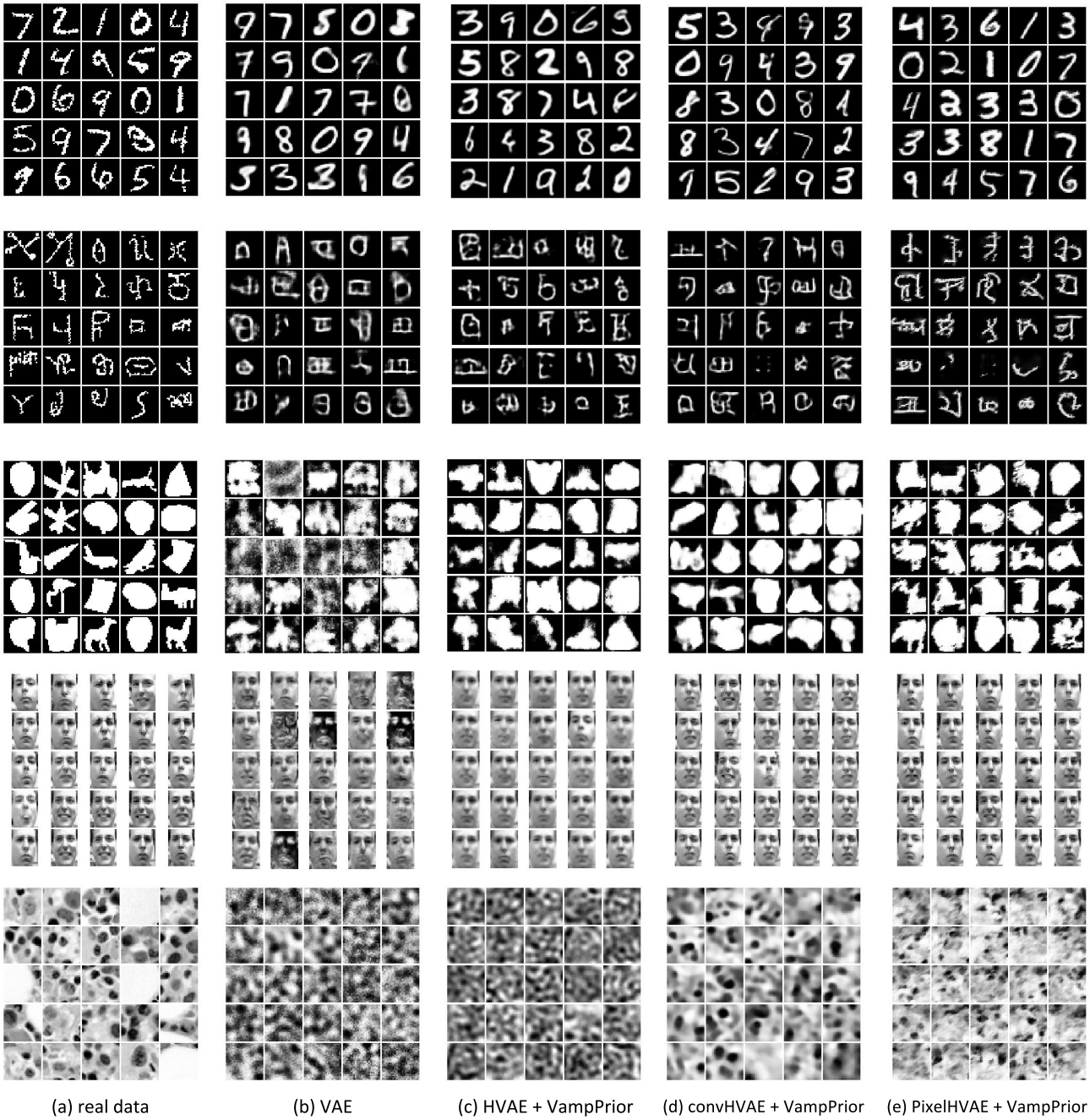

Figure 5: (a) Real images from test sets and images generated by (b) the vanilla VAE, (c) the HVAE ($L = 2$) + VampPrior, (d) the convHVAE ($L = 2$) + VampPrior and (e) the PixelHVAE ($L = 2$) + VampPrior.

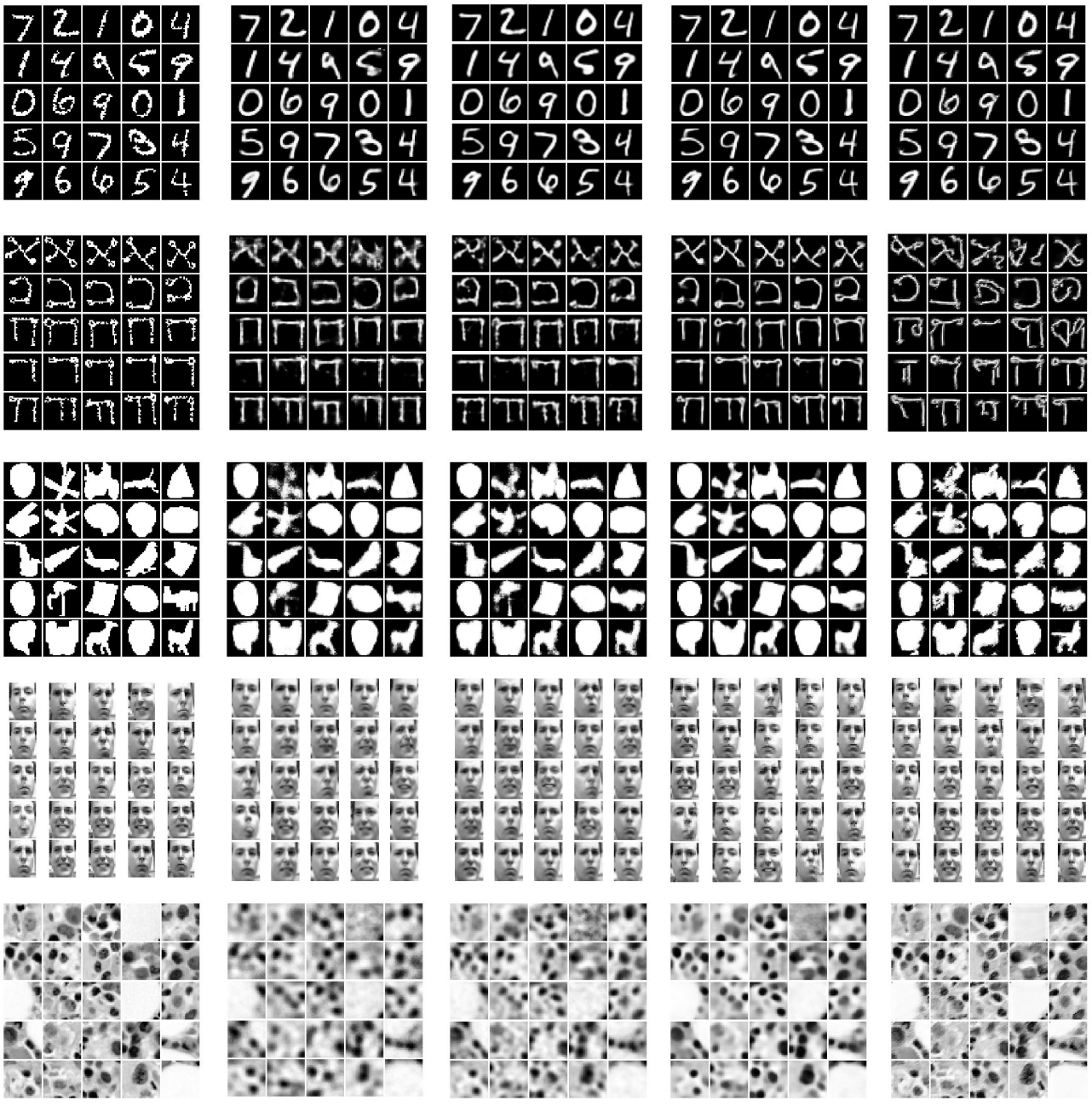

Figure 6: (a) Real images from test sets, (b) reconstructions given by the vanilla VAE, (c) the HVAE ($L = 2$) + VampPrior, (d) the convHVAE ($L = 2$) + VampPrior and (e) the PixelHVAE ($L = 2$) + VampPrior.

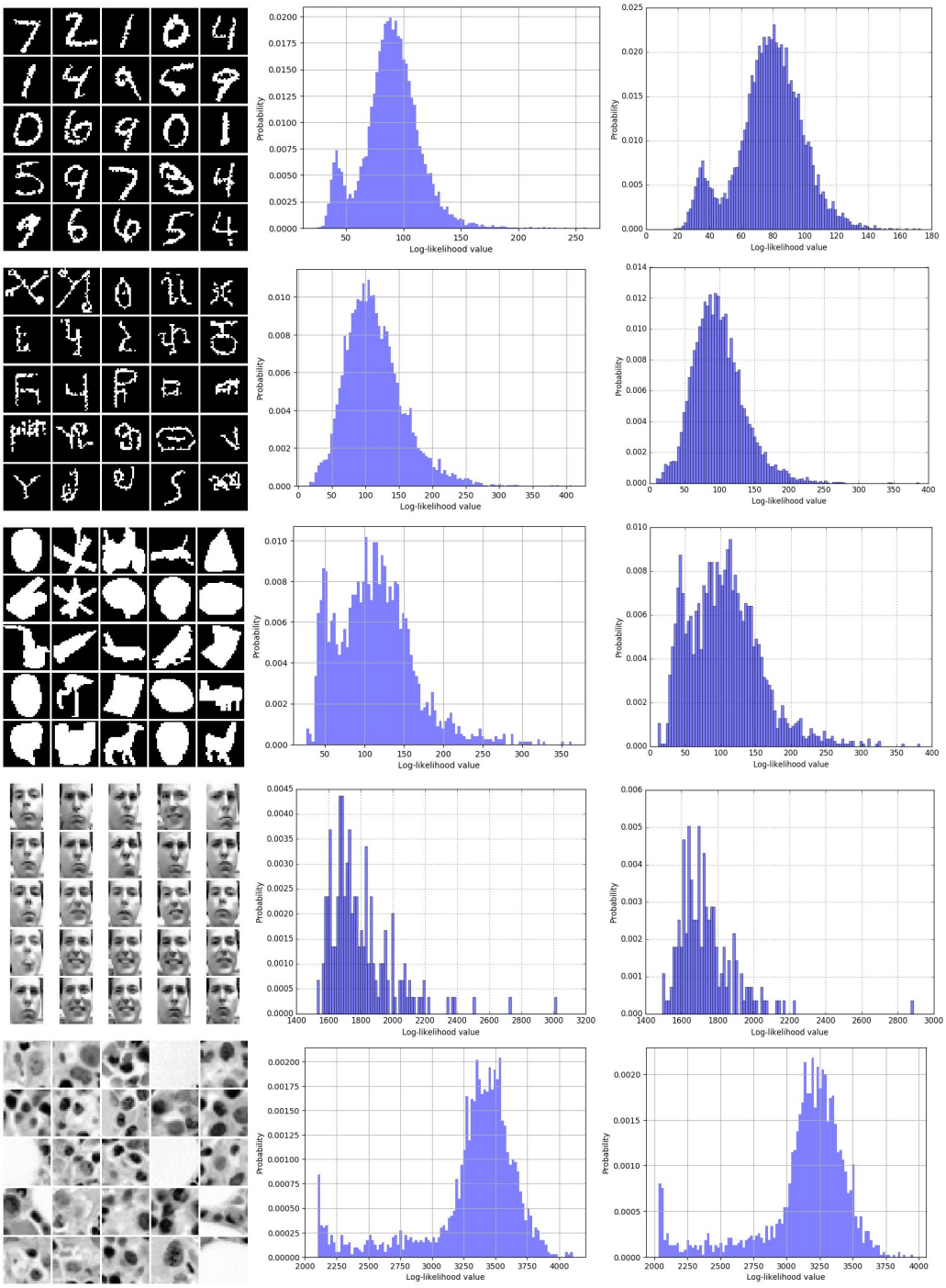

Figure 7: Histograms of test log-lihelihoods calculated on (from top to bottom) MNIST, OMNIGLOT, Caltech101Silhouettes, Frey Faces and Histopathology for (b) the vanilla VAE and (c) HVAE ($L = 2$) + VampPrior.